How DARPA's Dark Web search engine could help your business

[Katherine Noyes (/author/Katherine-Noyes/)](/author/Katherine-Noyes/)

IDG News Service   Feb 12, 2015 6:54 PM

A "Dark Web" search engine developed by U.S. defense researchers is in the spotlight this week for its use in combating human-trafficking activities, but it could play a role in business, too.

"There's huge potential," said Jeff Schneider, a research professor within the Robotics Institute at Carnegie Mellon's School of Computer Science. Earlier this year, Carnegie Mellon was awarded a $3.6 million contract to collaborate on the project at the Defense Advanced Research Projects Agency (DARPA), which is known as Memex (http://www.darpa.mil/Our_Work/I2O/Programs/Memex.aspx).

The Memex program—a three-year research initiative to develop software for domain-specific indexing of open, public Web content and domain-specific search capabilities—was highlighted on Sunday by a segment on 60 Minutes (http://www.cbsnews.com/news/new-search-engine-exposes-the-dark-web/) and an article in *Scientific American* (http://www.scientificamerican.com/article/human-traffickers-caught-on-hidden-internet/?page=1).

"The program right now is focused on human trafficking—that was chosen as the target domain," Schneider said. But the technology can be applied to other fields. "Any algorithms we can use to study human trafficking can easily be retargeted at other domains of people's interest."

We're all familiar with what's known as the "surface Web"—that set of Web pages that get accessed when we do searches through engines such as Google. The reason those pages are accessible is that they're connected to other public pages and are widely visible to search engine crawlers, also known as spiders.

What's different about the so-called Dark Web is that it's the set of Web pages that are not ordinarily accessible to Web crawlers because they're too fleeting, for example. By many estimates, mainstream search engines like Google index just 10 percent of what's actually on the Web; the rest are those pages that fall into the "dark" domain.

> **Further reading: Meet Darknet, the hidden, anonymous underbelly of the searchable Web** (http://www.pcworld.com/article/2046227/meet-darknet-the-hidden-anonymous-underbelly-of-the-searchable-web.html)

The potential of the "dark" side of the Web, however, could be considerable. Not only is it fodder for crime-fighting efforts such as the one being conducted with Memex, but it's also home to the majority of the electronic information out there.

Stock analysts might use Memex to search the Dark Web for information relevant to the stocks they cover, for example. Wine experts could use it to help them find the latest information in that realm.

"These technologies could be used to uncover fraud and other illicit activity in the business world," said Emily Kennedy, CEO of Marinus Analytics (http://www.marinusanalytics.com/), a Carnegie Mellon offshoot that's also involved in the Memex effort.

"Machine learning algorithms can help find patterns in massive amounts of data, and Memex takes that to an even deeper level," Kennedy said.
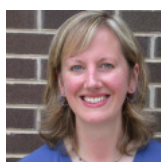
One of Memex's advantages in this area is its ability to poke around the Web and find pages not indexed by Google, Schneider said.

Another is its ability to tune its knowledge to specific domains of interest. With user feedback, it can keep refining those domains, he said.

"Users can tag pages as having certain properties of interest," he said, "and the tool learns those patterns."

Law enforcement already uses technology developed by Schneider's team, he said. Asked about what's coming next, he said, "this is a three-year research program that started six months ago. ... It's really just at the very beginning."

Related: Web & Social (/Category/Web-Social/)   Business (/Tag/Business/)

  Websites (/Tag/Websites/)

Katherine Noyes U.S. Correspondent

Follow @noyesk (https://plus.google.com/105492284418862345045/posts)